Visualizing structures of speech expressiveness

Herbelin, Bruno; Jensen, Karl Kristoffer; Graugaard, Lars

Published in:
Proc. Computers in Music Modeling and Retrieval

Publication date:
2008

Document Version
Publisher's PDF, also known as Version of record

Visualizing structures of speech expressiveness

Bruno Herbelin[1], Kristoffer Jensen[1], and Lars Graugaard[2]

[1] Aalborg University Esbjerg,
6700 Esbjerg, Denmark
{bh, krist}@aaue.dk
[2] Center of Design Research, Research Department,
Designskolen Kolding, Ågade 10, 6000 Kolding, Denmark
lhg@dskd.dk

Abstract. Speech is both beautiful and informative. In this work, a conceptual study going through the myth of the tower of Babel and hypothesis on the apparition of articulated language is undertaken in order to create an artistic work investigating the nature of speech. Our interpretation is that a system able to recognize archetypal phonemes through vowels and consonants could be used with several languages to extract the expressive part of speech independently from the meaning of words. A conversion of speech energy into visual particles that form complex visual structures provides us with a mean to represent this expressiveness of speech into a visual mode. This system is presented in an artwork whose scenario is inspired from various artistic and poetic works. The performance is presented at the Re:New festival in May 2008.

1 Introduction

In the speech process, the control of movements of specific parts of our body allows to control the generation of sounds to form words. This implicit transfer of energy from gestures into sounds provides voice with a great expressiveness. Speech and singing are archetypal and primary expression modes which, although languages have been developed in our societies mainly for communication purposes, tightly bind communication and expressiveness together. How to explore the possibilities of using the act of speaking in another way that for communicating words or, more specifically, to emphasize the expressiveness contained in oral communication?

For instance, the performance Ursonography [6] by Jaap Blonk and Golan Levin illustrates quite well how visuals can be automatically mapped to speech to enrich spectators' experience. However, in this case, the emphasize is on the text of Kurt Schwitters' *Ursonate* poem; the visual interpretations are given only through changes in the typography, but keep a direct link to the words themselves. The same limitation can be observed with the interactive poem generation proposed by Tosa et al. [12]. It is quite well known that visuals have a very deep impact in our perception of voice. Both perceptions are so tightly coupled in our brain that surprising links can occur; to mention only the most well known, we shall cite the McGurk effect [2]–where people think they heard 'DA' when hearing the sound 'BA' and seeing the lips pronounces 'GA' simultaneously– and the sound-induced flash illusion [11]–where a single flash accompanied by several auditory beeps are perceived as multiple flashes. Considering these limitations and the potential impact of visuals on the perception of speech, we propose to invert the process and exploit the richness of speech to generate visual feedback.

To illustrate and investigate this paradigm, we designed and implemented a system operating on voice to generate 3D visuals on the fly. The transfer of speech into a volumetric representation convert time into space, dynamics into shapes, eventually illustrating the duality of speech in an easily comprehensible way. However, this does not work with any action/reaction process nor a simple display of the audio wave signal. We propose to present how a process centered on archetypal structures of speech, on the contrary, meaningfully and expressively 'translate' speech into visual shapes.

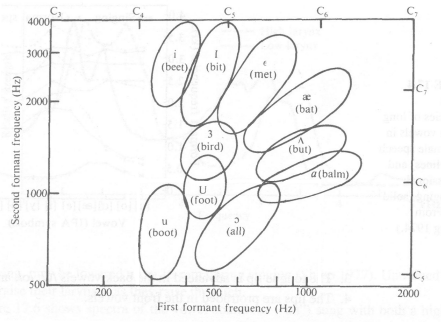
Section 2 details the speech articulatory gesture and establishes the basis for a 'phoneme-to-visuals' transformation by the introducing the concept of archetypal phonemes. Then, section 3 describes our approach to speech analysis and section 4 details the mechanism for the speech transformation and gives examples of the visual shapes obtained. This system and the basis for artistic performances are outlined in section 5, giving several examples of artistic uses. Finally, a conclusion is given on a more general level.

2 Speech and language

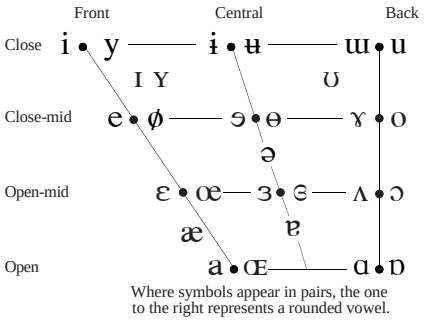
Speech is the physical act, by articulatory gestures in the mouth, throat and chest, of producing the voice. The airflow created by pressure from the lungs are, through obstruction of the vocal cords and/or mouth constrictions and correct positioning, shaping and moving of the tongue, jaw and lips, emitting sounds that are to be understood by the target persons. While communication is done by means of words and sentences, the underlying sounds are often used commonly in many words. These sounds are generally classified into two groups; vowels and consonants.

2.1 Vowels and consonants

Vowels are created by filtering the sound source through the shaping of the mouth cavity opening.The source consists of vocal cord vibrations or constrictive noises in the case of whispering, both sources are created using airflow generated by pressure from the diaphragm and the abdominal muscles. The filtering is generally understood as resonances, i.e. parts of the frequency spectrum that is increased. These resonances are called formants. The two first formants, F1 and F2, are often used in a formant diagram to illustrate the differences between the vowels. The formant diagrams for common vowels are shown in figure 1.a. Notice how some vowels overlap, meaning that the same F1 and F2 frequencies can signify different vowels.



(a) Formants diagram (After [10]).



(b) Internal Phonetic Alphabet of vowels [5].

Fig. 1. Vowel diagrams

The vowels are classified in the international phonetic alphabet (IPA) [5] through several articulatory features. The frontness describes the highest tongue position (front, central and back) and correlated with F2, and the height describes the jaw distance, generally in four steps, and correlates with F1. In addition to the frontness and height, the roundness (whether the lips are rounded or spread) affects the formants. The vowels can also be rhoticiced or nasalized and the tenseness can be changed. In figure 1.b the common vowels are shown as a function of the frontness (F2), height (F1) and roundness.

The consonants are classified in the IPA [5] by their place and manner (how a sound is produced from a vocal gesture). Constrictions or stops occur at various places in the mouth (points of closest constriction). The manners are, for instance, plosives (*p,k*), fricatives (*f,s*), or nasals (*m,n*). Much of the sound of the consonants is dependent on the neighboring vowels.

2.2 Archetypal phonemes

It is natural to consider that the phones are the basic elements of every language. Historically, though, every "village" contained some particularities (diacrites) of their phones that made it possible to recognize the identity of their original village by their neighboring villagers. More well known is the regional accents existing in most languages, sometime to the point of making comprehension impossible, or even, as in the case of American English, becoming another languages. This phenomenon is illustrated in the mythology with the story on Babel tower, built to reach the heavens; in the beginning "...the whole earth was of one language, and of one speech..." [3], but because of their disobedience, God confused their languages so that the men could not work together anymore. Still, languages are known to disappear by the day, and it may seem to be an inverse 'Babel tower phenomena', that distances in the global world 'shrink' to the point of allowing fewer languages.

In other words, if finding a universal identification of phonemes independently from the language is controversial from the linguistic point of view (because it is usually established that each language is based on its own set of phonemes, e.g. 62 in English, 32 in French or 28 in Italian), it has a philosophical and historical basis.

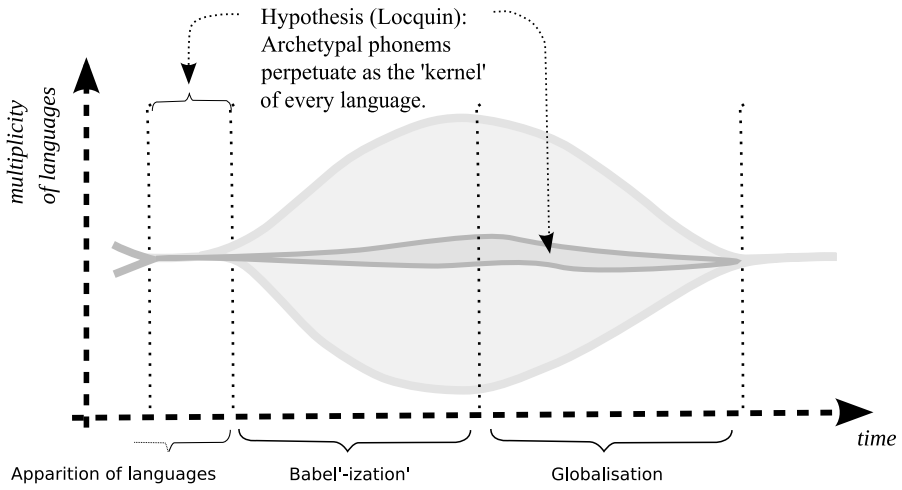


Fig. 2. Illustration of the multiplicity of languages after the supposed Babel effect, and the diminishing after the supposed globalization effect. Locquin has attempted to identify the archetypal phonemes at the beginning of languages.

Therefore, building a speech interface based on speech expressiveness rather than linguistic may be appropriate metaphorically, but would go against the natural tendency to distinguish and isolate languages. Here, the research work of Marcel Locquin may provide us with a good starting point. In his study on the apparition of articulated language in human pre-history [8], he defines twenty archetypal

[3] Genesis 11:1-9

phonemes in human languages which would have formed the roots of every current languages. These twenty phonemes were identified in elementary sounds of baby talk and in 80 extinct and 50 living languages[4]. Our approach differs from Locquin at this point as we do not need to investigate further into their meaning or their impact on the structure of languages, but we consider that the principle of a global classification of human 'emit-able' sounds into a limited number of classes offers interesting potentials for building a speech interface.

Figure 2 shows the motivation for our approach. Supposedly, in the beginning of language, there were fewer phonemes. The multiplicity of languages also increased the number of phonemes, but increased communication in the globalization age now decreases the number of phonemes. If Locquin could identify that there exists a limited number of archetypal phonemes, we may simplify the phoneme detection problem into a problem of identifying which of the chosen phonemes are actually pronounced.

2.3 Language uses

In this section, the common understanding of the reasons for working on language will be shortly given. As a common understanding, language is here supposed to be a filter that enables humans to access the reality of the world. However, the idea of an amodal cognitive function coupled with the multimodal sensory apparatus of the humans gives reason to doubt the uniqueness of language as a means of understanding the surrounding world. Instead, language is in part seen as a functional tool to increase the chances of survival. By detailed warnings of danger or possibilities of food, the language helped humans survive in difficult situations. Another important function of language is to ensure cohesion in a group. Huron [4] argues that the main purpose of the singing voice is to reach beyond the small group while bonding or performing collective actions –music being better adapted to larger groups than language, whose reach is 4-5 persons. Language is also important as to the mental representation of the world. Lucy [9] shows that different languages with different grammatical categories give rise to differences in attention, memory and classification. Language has thus a formative role on the human identity, and it can be hypothesized that a person can be 'metamorphosed' into another identity by changing its language. Furthermore, language is so essential for humans, to the point that Lacan calls humans 'Parlêtres' [1], thus putting language at the center of the human (un)consciousness.

These short observations on the use and limitations of languages are all referring back to the causes of the apparition of speech, to the reasons for its development, and to the ancestral and deeply human need for expressiveness when performing act of speech.

3 Detection of archetypal phonemes

Considering the potential universality of phonemes in articulated languages, the detection of phonemes in voice seemed to be an appropriate start for our attempts to transform speech into visuals. Existing software could not be used because of specific requirements differing from classical speech recognition:

- Events should be triggered as phonemes are pronounced, not after a complete analysis of a word or a sentence.
- There is no need to identify words for their meaning in a dictionary.
- Vocal attributes such as amplitude and speed may affect the output.

Therefore, we had to develop a custom system providing these features in real time. Our program implemented using Cycling 74's Max/MSP [5] is described below.

[4] http://trans-science.cybernetique.info/fr/archetyp.htm, accessed 20 March 2008.
[5] http://www.cycling74.com/

3.1 Principle

We analyze for perceptually significant data of noisiness and bark scale, as well as pitch and spectral centroid. The FFT analysis uses a analysis window of 2048 samples with a 512 samples overlap, giving a latency of approximately 21 msec. We use a 25 band bark scale decomposition of the spectrum for storage and comparison of vowels and pitched consonants. A noisiness parameter is used to determine certain fricatives.

The user stores a snapshot of each of the vowels and pitched fricatives to be detected. The comparison is done in real-time by identifying the distance between the input bark vector and each of the stored vectors. The resulting distance vector – its length equal to the number of stored vectors – is then searched for the shortest distance. We identify a specific vowel (or a specific pithed consonant) when the shortest distance is at the same index in three consecutive vectors, and when the noisiness is below the fricative threshold. We identify a 'fricative state' when noisiness is above the fricative threshold and the loudness is above the background noise level.

3.2 Vowels and Consonants

The output of each bark channel is low-pass filtered and normalized over five analysis windows. Stability of vowel detection is obtained by requiring three successive detection positives before passing a true. The shortest latency is therefore equal to three window overlaps, or approximately 35 msec.

Fricatives found in e.g. plosives are very short in duration, and therefore not possible to detect with the vowel detection technique, even though significantly different on the bark scale. We have therefore opted for detecting fricative/non-fricative, where increased noisiness and vowel instability determines the presence of a fricative. To avoid false positives when background noise is primary input, we have set a loudness threshold before fricative detection kicks in.

Pitched consonants have just as significant bark vectors as vowels. Since they also have long enough duration, their detection can be obtained by using the same technique as the vowels.

Phoneme detection of sung text is rather more complex, because the pitch assigned to a vowel affects their articulation. In musical composition this is a well-known fact, and it means that the same vowel may have a significantly different bark vector at different moments of the performance. This can be compensated for in the composition process by good choise of pitches for each vowel, but this may contradict the artistic and musical aims at a given moment in the composition.

In addition to the parameters used for spoken speech, we have incorporated pitch stability measured as the stability of a fundamental pitch over the current time frame.

4 Expressive visualization of speech

In this section, we propose to explain our approach for transferring speech into visuals. As our intention was not to reproduce a vocal interface based on the meaning of words, we had to go back to the elementary elements forming speech, the phonemes.

4.1 Particles of speech

To produce a visual feedback to speech, we imagined that each phoneme in the speech would be sent into space as a particle. This phenomenon can be compared to the physical model of photons used to describe light, but applied here to what we call 'phonemons' to describe speech. Concretely, we wanted to build an interactive process capable of showing these 'phonemons' getting out of the mouth when someone speaks.

Using this equivalence between phonemes and particles, the transfer of speech into visuals appears to be simple, which is usually a good criteria for interaction. The other advantage is the one-to-one mapping between modalities which should support the reproducibility of the phenomenon, thus enabling an apparent control of the process by the user.

4.2 Energy transfer

The live detection of phonemes can ensure that a particle of speech (a 'phonemon') is sent in space each time a phoneme is recognized. To describe how these particles of phoneme evolve into space afterwards, we attributed the pseudo-physical behaviors of a spring-mass particle system to the 'phonemons', so that their energy would mirror the attributes of voice (loudness, frequency, etc). However, in a spring-mass systems, particles either converge in an equilibrium or diverge in a chaotic explosion. To avoid the latter, we fine tuned the parameters to guarantee a convergence in a controllable delay, so that spectators would always see the full transformation process of speech into stabilized spatial structures.

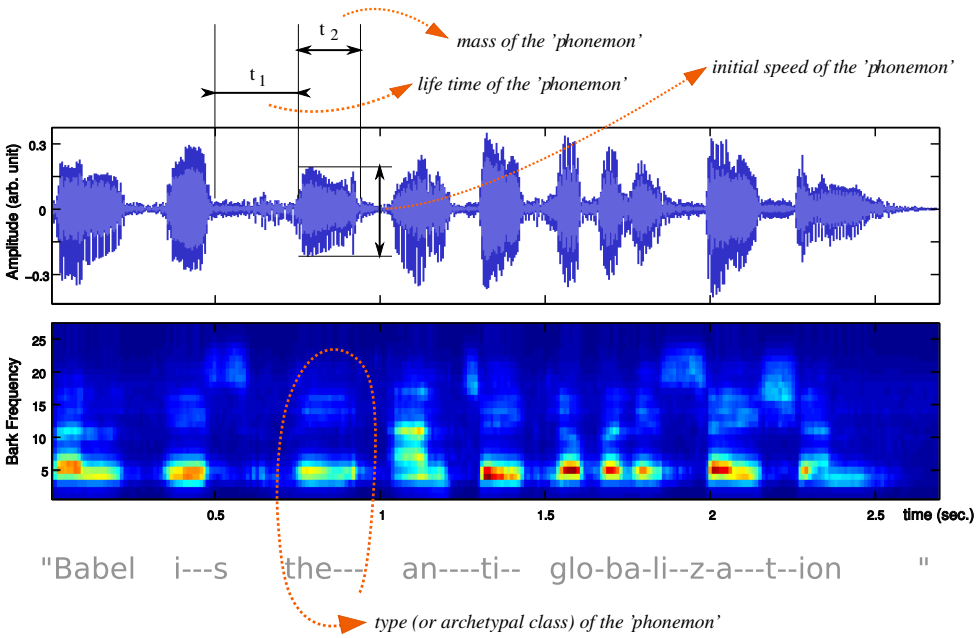


Fig. 3. From sound to particles; transfer of energy from signal analysis into pseudo-physical attributes of phoneme particles.

The main parameters for particles are their mass and initial speed when sent into space. One other parameter of interest for the process is the life time of each phoneme which controls the formerly mentioned delay of stabilization. The transfer of energy of voice into particles is summarized in figure 3 and below:

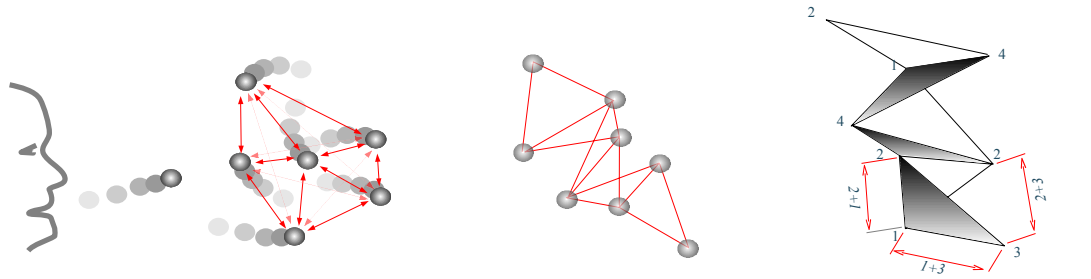
– The mass of a particle is given according to the the duration of the phoneme, because it reflects the importance (weight) of a phoneme in a word. For the singing voice, this is weighted with the pitch stability.
– A particle's emitting speed is given according to the amplitude of the voice in order to reflect the intensity of speech (wispers go slowly and not far, shouts go fast and far).
– The life-time is given according to the delay with the previous phoneme to evoke the dynamic of speech (slow speech takes longer to stabilize).

Visually, particles are points which size varies according to their mass. From the user point of view, the louder he/she speaks, the faster the particles will go, and the slower he/she speaks, the heavier the particles will be and the longer they will live.

Up to there, only prosodic attributes of speech are captured (like duration or amplitude of phonemes). In order to capture the acoustic variations in speech, we attribute a type to each particles according to the recognition of the phoneme. As the potential existence of archetypal phonemes suggested by Locquin supported the possibility to classify phonemes in a limited number of archetypes (see section 2.2), we used a limited set of particle types (ten or twenty, but this could vary). This last addition was visually represented by a different appearance and color for each particle type.

4.3 Modeling shapes from speech

Thanks to the spring-mass animation of particles of phoneme, we eventually transform speech into a cloud of points floating into space (figure 4.a). Connecting these points to form polygonal shapes appeared to be a natural continuation; for instance, a word made of three phonemes would lead to a triangle, a word of four phonemes to a tetrahedron, and so on to form a ribbon of potentially several particles (figure 4.b).



(a) A particle is emitted for each phoneme, and then interacts in a spring-mass system.

(b) All particles of a sentence form a ribbon in space.

(c) The shape generated depicts the original words pronounced.

Fig. 4. Phonemes extracted from speech interact in a cloud of particles before building a mesh. The energy transferred from voice to particles is kept and the archetypal phoneme type (number) is used in the last phase of stabilization.

In order to ensure the reproducibility of the modeling of shapes from voice, we use the type attribute of the particles to influence the relative distance between them. This ensures that the position of vertices in a bloc is computed according to the original voice analysis (i.e. the length of the 'spring' between two particles is the sum of the lengths for each type). This way, two blocs of the same length would become two structures with the same topology (same number of particles) but with a different shape (each particle being different). Figure 4.c gives an example of the structure built from particles of different types.

In addition, a detection of silences was performed on voice to isolate parts of speech, as words or sentences. This simple but essential process allows to put an end to the building of shapes coming from a continuous and potentially infinite flow of words: all the phonemes between two silences should belong to the same chained structure that we call a 'bloc' of speech. Of course, the length of a silence is a subjective criteria which is unrelated to a language-specific knowledge of words –it would rather coincide with the silence between two sentences or when the speaker take a breath. The duration of this delay between sentences was subjectively adjusted later on (approximately one second).

At this stage, a 'bloc' is a mesh made by triangles striped according to tetrahedrons of four consecutive vertices (particles), very much like a ribbon folded in and out. From the graphical point of view, this leads to the construction of a chain of tetrahedrons (four particles) with common faces (three former particles) and sized according to the weight of the vertex (types of phonemes). Figure 5 shows that the structure is really dependent on the succession of phoneme types; when providing algorithmically regular sequences, we obtain regular shapes (alternative cycle in figure 5.a. or growing and shrinking sequence in figure 5.b.) and different sentences pronounced produce different shapes (figures 5.c. and 5.d. –you may notice the glow around vertices added to enhance the impression of light).

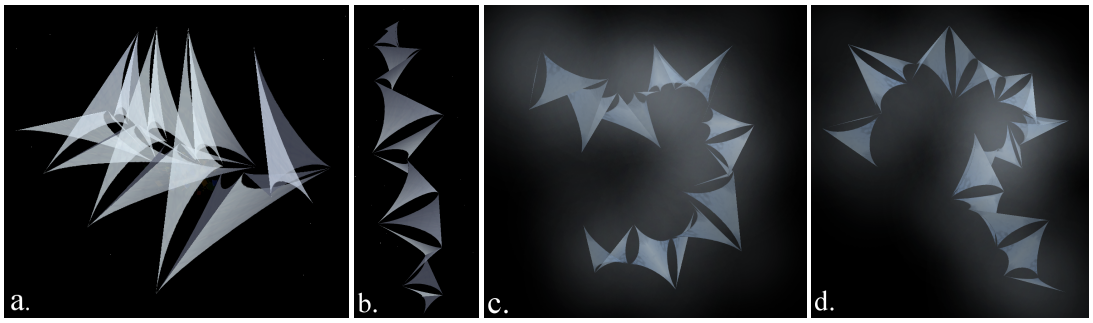


Fig. 5. Examples of the 3D representation of various blocs of speech; two regular structures generated artificially (a. is the alternations of types 1/10, while b. is the linearly changing chain of particles type 1-2-3...-9-10-9...-3-2-1) and two irregular shapes actually from speech (c. and d.).

To sum up, the original idea was to map the output of a phoneme detection (amplitude, duration, time since last phoneme, silences and phoneme type) to both dynamic and structural parameters of the visualization, eventually influencing the spatial organization of vertices, thus forming 'blocs of speech' representing the spoken sentences. It is important to notice that, at this stage of our voice-to-3D transformation, phonemes' durations and types have been preserved and integrated inside the geometric structures as masses and distances. To the opposite, voice amplitude and speed were transformed into dynamic factors only –the speed of particles being anyway transferred to the blocs in the form of kinetic energy. The reason for this choice is that the duration and the former are part of the linguistic structure identifying the message, whereas the latter can vary very much from one pronunciation to another.

5 Experiments

The 'translation' of speech into 3D structures using particles as elementary graphical elements has been used in various artistic contexts. Here are the main experiments and performances presented.

5.1 The original interactive installation

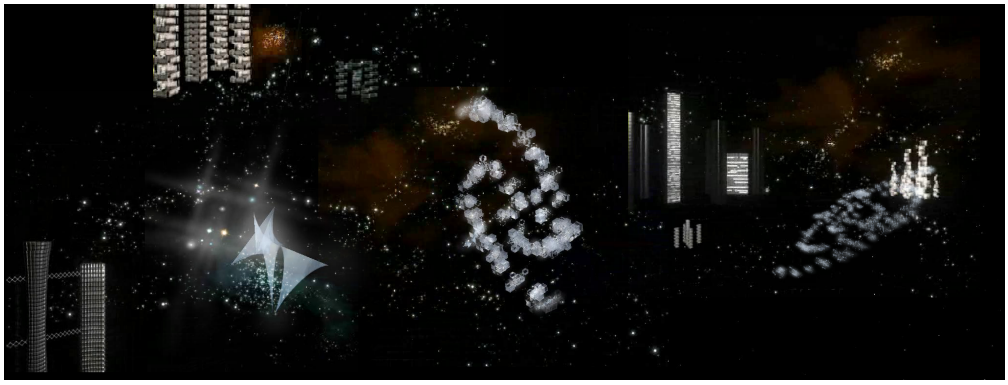


Fig. 6. Reconstitution of the stellar landscape where particles (bright points) are sent into space to form the architectural structures of the Flying Cities.

Originally presented in 2003, the "Flying Cities" project [3] allowed people from three countries (France, Italy and Germany) to interact seamlessly in an interactive installation by using their own language. The artistic aim of this creation was oriented towards the production of artificial architectures from people's speech, and was the context of our original experiments on the transformation of speech into 3D structures. The materialization of the elementary phoneme's acoustic attributes into forms is also a metaphor for civilization's expressions of itself into architectural forms, because language and beliefs construct word-built civilizations (figure 6). As such, the project indirectly refers to the 'Babelization' process in the evolution of languages (Section 2.2). This project was also inspired from the work of Russian constructivist artist, and more specifically by Klebnikov's experiments on his poetic language 'zaum', a 'stellar language' based on elementary phonetic structures [7].

Based on the observation of people's reactions, we could notice that the speech-to-3D mapping established for this experiment could generate new sensory experiences of language as architecture, and materialization of sound and pattern as spatial forms. The interaction paradigm used for Flying Cities succeeded in showing the interrelationships of multiple modalities, including voice and words within a visual and musical environment. The conceptual construction of the visual conversion combining both structural rules and artistic interpretation provided understandable correspondences and complementarities between the audio and visual senses.

The installation originally developed didn't however give the possibility of exploring the generated virtual architectures. This was observed as an important limitation for the apprehension of the whole processl; this could be improved with a better control of the visualization. We believe that the perceptual process of associating a visual result with a vocal input is possibly sufficient in itself not to require an important scenographic support as originally supposed.

5.2 Musical experimentation

To experiment further with this principle of particles for sound, we also made a rapid prototype of the possible use of our visualization system to music. We replaced the voice by a MIDI saxophone and produced a particle for each note. The equivalence to the prosodic and structural elements of speech to

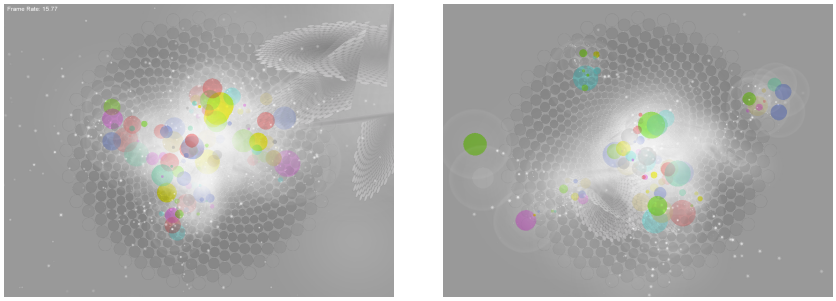


Fig. 7. Visualization of musical performance using the same principle; each particles is a note with color, size and other attributes mapped to the MIDI signal (instead of phonemes for the speech version).

music was simple to do (MIDI note, pitch, volume, etc.) and the system was presented at the opening banquet of the ICAT 2008 conference[6]. Figure 7 shows the visual adaptation we made for the occasion.

Music performances accompanied by visuals are nowadays very common, but the generative power and dynamics of the particles produced pleasing results. Nevertheless, the originality of the approach did not compensate for the repetitiveness of the process; the cause was mainly in the impossibility for the musician to exclusively play with generating particles as he also had to play some music! We could however gain experience and knowledge from this experiment. First, the easy adaptation to music proves that our audio-visual mapping is language independant, and gives a first insight on how to explore further the musicality of speech. This allowed us to improve the implementation of the visuals, like in the way particles appear during the production of a long sound, or in the elaboration of extra dynamic effects. Finally, this gave us the confirmation that the process could be used for a performance instead of an installation.

5.3 Current Artistic aim

A new experimentation is scheduled to take place at the Re:New 2008. It consists of a performance involving a singer and a screen. While the performer sings (poems and original text), visual particles are emitted on the screen as the speech is pronounced. Three possibilities exists here, if the particles has enough energy, they will attain and form a central structure, while if they do not possess enough energy, the particles will not reach the central structure, but instead either fall to the ground or disappear.

The interpretation of the central structure is left open, with several possibilities, for instance, that it is related to the Tower of Babel, and all the speech is found in there. As in the myth, all knowledge is accumulated in the heaven (knowledge). Unfortunately, the immensity of the tower of Babel make communication impossible, and all the particles regrouped in the tower cannot understand each other anymore. While the speech particles are more and more easily attaining the tower, the process shows the incomprehension and lack of accumulation of signification. Many languages will be spoken during the performance (Danish, English, French, Italian, ...) which will add to the incomprehension.

Another interpretation is that the central structure is a being that is taken shape as the speech particles are reaching and creating it. As more and more particles are regrouped, the central structured shape gets to a more active state and its skin is formed in order to show the structure as one whole. As the structure is getting alive, a dialog can take place between the performer and the structure. This dialog is in the form of speech that is returned to the actor, but also speech particles that are emitted from the structure to the actor, or to form new structures in the space.

[6] The Danish band DuoResonant was playing. See http://www.icat2007.org for details on the conference.

Finally, the interpretation can be that the structure is a metamorphose of the actor, and eventually, the actor is disappearing when the structure is forming a being from him/her inner matter. In this situation, new speech particles is also affecting the already formed blocks, pushing and disturbing them, until thee are big enough to withstand the disturbances.

6 Conclusion

Speech expressiveness is understood in this work as the non-linguistic communication part of vocal activity. As a starting point to this discussion, speech was described as the continuous concatenation of vowels and consonants. Vowels and consonants are fully described in the Internal Phonetic Alphabet as the basis for several hundred possible different phonemes. To support our simplifying assumption, we presented a supposed evolution of this large variety of phonemes as being originated from a limited number of archetypal phonemes (as supported by Locquin) which would have evolved and multiply in the 'Babelization' of human languages while remaining at the core of current languages.

Based on the joint principles of universal identification of phonemes in speech and cross-modal energy transfer, we have designed and implemented a system extracting and converting the structures of speech into visual representations. This process was designed to be consistent (an action always produces the same outcome), invariant (a kind of conservation of energy shall apply from inputs to outputs), and generative (combining actions allows to build something). It was supported by elementary principles like the real time visualization of 'speech particles' sent into space as phonemes are detected in speech, the linear mapping of amplitude or duration into parameters of a spring-mass system, or the interpretation of the phonemes' type into construction rules for the resulting 3D structures.

The main qualities of the proposed transfer of speech into 3D shapes are the easiness to be apprehended and understood by spectators and its ability to effectively transfer the expressiveness of speech while totally ignoring the message contained in the words. Such speech interface presenting a visual analogy of the transfer of energy being emitted by humans as speech patterns has the power to be meaningful for people thanks to the references made to the scientific, linguistic, and artistic fields.

References

1. Jean-Michel Rabaté (editor). *The Cambridge companion to Lacan*. Cambridge University press, 2003.
2. McGurk H. and MacDonald J. Hearing lips and seeing voices. *Nature*, 264:746–748, 1976. http://www.media.uio.no/personer/arntm/McGurk_english.html.
3. B. Herbelin, S. Lasserre, and J. Ciger. Flying cities: building a 3d world from vocal input. *Journal of Digital Creativity*, 19(1):62–72, January 2008.
4. David Huron. Is music an evolutionary adaptation? In Robert Zatorre and Isabelle Peretz, editors, *The biological foundations of music*, pages 43–61. Annals of the New York Academy of Sciences, 2001.
5. International Phonetic Association. International Phonetic Alphabet (IPA). http://www.arts.gla.ac.uk/ipa/ipachart.html, 2005. Accessed on Jan. 23 2007.
6. Blonk J. and Levin G. Ursonography. Ars Electronica Festival. Linz, Austria, April 2005. http://flong.com/projects/ursonography/.
7. Velimir Khlebnikov. *Des nombres et des lettres (essais reunis)*. L'Age d'Homme, 1986.
8. M. Locquin and V. Zartarian. *Quelle langue parlaient nos ancêtres préhistoriques?* Albin Michel, Paris, 2002.
9. John A. Lucy. *Grammatical Categories and Cognition. A Case Study of the Linguistic Relativity Hypothesis*. Cambridge University press, 1992.
10. G. E. Peterson and H. L. Barney. Control methods used in a study of the vowels. *J. Acoust. Soc. Am.*, 24:175–184, 1952.
11. Ladan Shams, Wei Ji Ma, and Ulrik Beierholm. Sound-induced flash illusion as an optimal percept. *Neuroreport*, 16(17):1923–1927, November 2005.
12. Naoko Tosa and Ryohei Nakatsu. Iteractive poem system. In *Proceedings of the sixth ACM international conference on Multimedia*, pages 115–118. ACM, 1998.